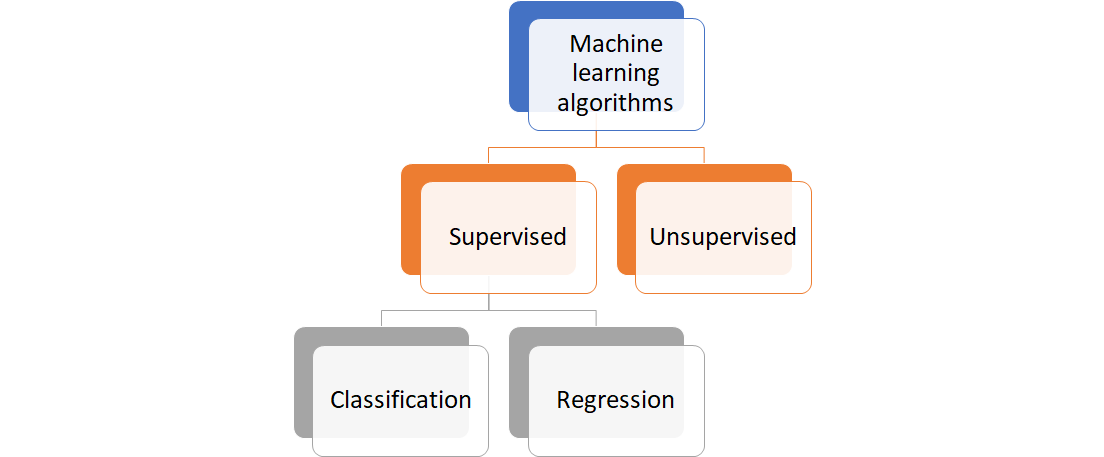
# The basics of Machine Learning – short lectures

In addition to ***chapter 4 of the reader ‘Predictive Modelling’***, this document will give an explanation of basic machine learning terminology. For most topics a short lecture is made (a movie of ~5 minutes) that introduces the topics. An overview of the short lectures can be found below (click on the link to go to the Hanze mediaspace):

1. [Classification](https://video.hanze.nl/media/MLA+Classification+-+Quiz/0_g6a4e62d)
2. [Evaluation of classification models](https://video.hanze.nl/media/MLA+Classification+Evaluation+-+Quiz/0_3kkxevtc)
3. [Regression (and evaluation of regression models)](https://video.hanze.nl/media/MLA+Regression+-+Quiz/0_hvdbo5he)
4. [Over- and underfitting with supervised models](https://video.hanze.nl/media/MLA+Over-+and+underfitting+-+Quiz/0_nlz4wgzj)

# Machine learning

The basics of the machine learning landscape are depicted in the figure on the right. The difference between supervised and unsupervised learning is discussed below (‘Supervised versus unsupervised’). The short lectures explain how to train supervised learning models. For both classification and regression it is discussed what these models are, how to evaluate them, and how to detect and prevent over- and underfitting using a training and a test set in order to create a good model that works well on new data. But first, let’s look closely at the difference between supervised and unsupervised learning.

## Supervised versus unsupervised

Machine learning tasks can generally be grouped into two main sub-categories: *supervised* and *unsupervised* learning. They are used for different types of problems. The difference is that with supervised learning, the actual outputs (or: *labels*) are known while training the algorithms, while with unsupervised learning this is not known. Unsupervised learning algorithms are generally used to detect some sort of structure in the data. Below examples of both types of learning are discussed.

## Supervised learning

In supervised learning, the class labels in the dataset which is used to build the classification model, are known. For example, a labelled dataset for flower-species would contain observed example data about some characteristics together with the actual label of the species. In a supervised learning problem, we would know which kind of flower in the training set is associated with which kind of characteristics. We use this information to train our model in order to classify new unseen flowers based on measured characteristics.

As another example, say we scrape the data from funda.nl, a well-known Dutch site for houses that are for sale. Here we can find all sort of information about houses: the size of the house, of the lot, the number of bedrooms, the asking price etc. Say we also have access to the data from the Kadaster, the Land Registry of the Netherlands, so that we know for each house that was sold for which price it was sold.

We can now make a predictive model where we use all properties of houses that were sold as features, and all selling prices from the Kadaster as the labels. We train a model that can predict these selling prices: this model will learn what features are important to make a good prediction for the selling price. This model can learn this because the labels – the selling prices – are known. When a new house appears on funda.nl, we can use our model to predict what the selling price for this house might be based on its properties.

Getting the labels can be a problem. Sometimes when the labels are not known, they have to be created by hand (for example, when classifying images).

Supervised learning can be divided into regression and classification problems. This is discussed in the short lectures, together with how to evaluate regression and classification models:

* [Classification](https://video.hanze.nl/media/MLA+Classification+-+Quiz/0_g6a4e62d)
* [Evaluation of classification models](https://video.hanze.nl/media/MLA+Classification+Evaluation+-+Quiz/0_3kkxevtc)
* [Regression (and evaluation of regression models)](https://video.hanze.nl/media/MLA+Regression+-+Quiz/0_hvdbo5he)
* [Over- and underfitting with supervised models](https://video.hanze.nl/media/MLA+Over-+and+underfitting+-+Quiz/0_nlz4wgzj)

## Unsupervised learning

In contrast to supervised learning, unsupervised learning deals with unlabelled instances. In unsupervised learning classes have to be inferred from the unstructured dataset. For example, continuing with the housing dataset discussed above, say we only have the data from funda.nl but we do **not** have the actual selling prices of sold houses. Now we cannot make good predictive models to predict the selling prices, but we could investigate our dataset: for example, are there groups in our dataset of similar houses? This is called *clustering*. Clustering is about ordering data into groups based on certain similarity measures: the algorithm will find by itself what items are similar. Maybe if we can find different clusters in our dataset, we can use this to learn about asking prices. If a new house lies in a cluster because it has similar properties as the other houses in a cluster, this can help to determine an appropriate asking price for the house.

## And more…

There are other types of learning algorithms, such as reinforcement learning. But this is out of scope for this class, and hence not considered here.